# Slovak Morphosyntactic Tagset

*Radovan Garabík and Mária Šimková*
Ľ. Štúr Institute of Linguistics of Slovak Academy of Sciences,
Bratislava, Slovakia

## ABSTRACT

Morphological annotation constitutes essential, very useful and very common linguistic information presented in corpora, especially for highly inflectional languages. The morphological tagset used in the Slovak National Corpus has been designed with several goals in mind – the tags are compact and easily human-readable, without sacrificing their informational contents. The tags consist of ASCII letters, numbers and several other characters. In general, they have a variable number of symbols, but their order is obligatory, and each category or specific feature is assigned a particular character, which can be shared among several parts of speech. The tagset is highly functional and pragmatic, although some allowances had to be made to accommodate the traditional analysis of Slovak morphology and part of speech categories.

*Keywords:*
*Slovak language, corpus, tagset, morphology, part of speech, grammatical categories*

## 1 INTRODUCTION

Morphological annotation constitutes fundamental and very common linguistic information found in corpora, especially for inflectional languages. It comprises the part of speech categorisation of lemmas and morphological characterisation of a word (token).

It is usually preceded by the process of lemmatisation (an assignment of the basic form to a particular lexeme). Since Slovak belongs to a family of highly inflectional languages, a morphological annotation is not a simple and straightforward process. Currently, the process of morphological analysis of such languages is often performed in two steps; the first one is the analysis itself (assigning to each of the words a list of possible combinations of lemmas and morphological tags), and

the second one is disambiguation, picking up one (correct, if possible) lemma-tag combination. The analysis itself is often nothing other than selecting the entries from a database of inflected wordforms (with an additional step of guessing lemmas and/or tags for out-of-dictionary words). The second step is often performed using statistical methods, requiring training on manually annotated corpora.

In the Slovak National Corpus (SNK), morphological annotation and lemmatisation occurs prominently in two places:

- manual morphological annotation and lemmatisation of the *r-mak* subcorpus
- automatic morphological annotation of the whole corpus (and other relevant corpora and subcorpora)

The *r-mak* subcorpus is a manually lemmatised and annotaded corpus of 1.2 million tokens (punctuation included). The progress from version 3.0 (released in 2008) to 4.0 (released in 2012) did not encompass any new texts; rather the existing annotations have been semi-automatically proofread and corrected, several duplicities have been identified and removed, the revision of the tagset has been applied where necessary, and a new, more consistent sentence segmentation has been introduced.

Thus, the end users of the corpus (corpora) meet the analysis while using the corpus, either when entering more complex queries or when displaying grammatical categories of the results. In this article, we describe the tagset in detail, including the motivation behind some design choices.

As the Slovak National Corpus at its inception in 2002 was primarily aimed at linguistic (mainly lexicographical) use, the morphological annotation and tagset were created with this in mind – the design of the tagset was based on the formalised Slovak language morphology (Páleš, 1994; Benko *et al.*, 1998), traditional grammar description (Dvonč *et al.*, 1966) and other similar tagsets of related inflectional languages (Hajič and Vidová-Hladká, 1997; Džeroski *et al.*, 2000; Hajič, 2000; Dębowski, 2001). Tokenisation, lemmatisation and the principles of morphological annotation used in manual tagging of the *r-mak* corpus are described in the user guide (Garabík *et al.*, 2004). The tagset is used in the morphological database of the Slovak language,

covering (at the time of writing) more than 97 thousand lemmas and about 3.2 million inflected and tagged entries (Garabík, 2006).

The new revision of the tagset and some of the principles occurred in 2012. It did not introduce any new tags, but rather clarified many borderlike cases and the classification of many words has been re-evaluated (based on actual corpus evidence and inconsistencies introduced therein). This article presents some of the reasoning behind the decisions.

All the examples used in this article are based on actual text occurrences in the Slovak National Corpus.

## 2 TOKENISATION

The tagset is designed to cover morphology of the smallest possible units – this governs the tokenisation principles. Most notably, there are no multi-word tokens; each constituent of such an element is a separate token. This includes also hyphenated words – expressions like *slovensko-poľský* (Slovak-Polish) will be tokenised as three tokens: *slovensko*, - (i. e. the hyphen), *poľský*. The advantage of this is a clear and unambiguous approach to the tokenisation, but as a main disadvantage, we lose a reasonable way of dealing with multiword expressions, and even have to introduce a special morphology tag to mark constituents of such expressions.

## 3 OTHER SLOVAK LANGUAGE TAGSETS IN USE

### 3.1 *Tagset developed at the Institute of Formal and Applied Linguistics*

This tagset is an adaptation of the Czech language tagset developed at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague (Hajič, 2004). It is a positional tagset of fixed length, each tag containing 15 ASCII characters. Each position encodes one (grammar) category, and some of the positions are empty (13, 14). The first position encodes part of speech.

A distinguishing feature of this tagset is a very detailed description of the part of speech subdivision (position 2): e.g. there are 16 different types of numerals, 21 types of pronouns.

Most notably, the tagset does not encode verbal aspect (an omission inherited from the Czech tagset). With some effort and a database of perfective and imperfective verbs, it can be inferred from the lemma – indeed, for the Czech language this has been done in the extended version used in the Czech National Corpus[1].

### 3.2 *Majka/Ajka*

*Majka* is a Czech language morphological analyser developed at the Faculty of Informatics, Masaryk University in Brno (Šmerk, 2010), a reimplementation of the previous analyser *ajka* (Sedláček, 2001). The tagset has been carried over to the Slovak version of *ajka*. It is an attributive tagset, with one-letter codes for the attributes and one-letter codes for the values. The codes for the values can be reused across attributes – the tags are of unequal length (although a rather important feature is that the value assignment does not depend on a part of speech).

### 3.3 *Multext East*

The EC INCO-Copernicus project MULTEXT-East Multilingual Text Tools and Corpora for Central and Eastern European Languages (Dimitrova *et al.*, 1998) developed language resources for six Central and Eastern European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, as well as English.

Slovak language morphology specification compatible with the MULTEXT-East (MTE) tagset was not part of the original Multext East specification – it has been developed separately at the Ľ. Štúr Institute of Linguistics (Garabík, 2011). The tagset follows MTE principles and tries to be compatible with the other MTE language tagsets, and especially with Czech (some of the design features were directly inspired by solutions deployed in the Czech MTE tagset). The tagset has been influenced by the Slovak National Corpus tagset described herein – one of the design goals was to make an automatic conversion from the SNK tagset into the MTE not too difficult. This even meant removing some features from the MTE tagset if they could not be inferred from the information about SNK tag and lemma (e.g. the verb *byť* (to be) is always referred to as Type=c(copula)).

[1] `http://korpus.cz`

This tagset is used exclusively in the scope of the MTE project and related research (Garabík *et al.*, 2009).

## 4 GENERAL PRINCIPLES

While the attributive versus positional tagset dispute is not a very important one (after all, a tagset is just a representation of grammar features and a mapping from one representation into a different one provides no inherent insights), we have to realise that a tagset is something that will be with the users for some time, especially if we are designing a tagset to be used in a big ('national') corpus. Both attributive and positional tagsets have their advantages and disadvantages – the positional system is often opaque to the user; if the number of positions is big enough, it is difficult to find out which position is which without counting the positions. Attributive tagsets tend to be even longer, because each value has to be accompanied by its attribute; but if the attribute abbreviations are selected sensibly, the users can decode the meaning at a glance.

In designing our tagset, we tried to extract the best of the two words while keeping the disadvantages to a minimum. One of the most important design decisions is to keep the meaning of codes unambiguous – one letter should correspond to one value *only*, even across parts of speech. The only exception is the paradigm category, which reuses the part-of-speech code. We try to assign mnemonic, easily-remembered codes familiar from common Slovak education and linguistic environment whenever possible. The tags are of unequal length, but most tags follow the same structure for the same inflectional paradigm (not the part of speech category). These principles make it easy to test grammar categories in software. Checking the part of speech category could be expressed in a Python-like pseudolanguage as[2]:

```
if tag[0]=='S': # noun
    # proceed with the noun
```

and checking for the value of the grammar category can be as easy as:

[2] We are counting the positions from zero.

```
if '4' in tag: # accusative
    # proceed with the accusative
```

### 4.1 *Tag Structure*

As the part of speech information is often the most important, it is encoded in the first[3] position. The second position usually marks an inflectional paradigm for words that do have this category. The code for the position repeats that of corresponding parts of speech, e.g. `SS`… stands for a noun with a noun-like inflectional paradigm, and `PS`… for a pronoun with a noun-like inflectional paradigm. First we describe the symbols and corresponding grammar categories, then we discuss motivation and principles behind several choices for individual part of speech categories.

The tag can be optionally followed by a marker separated by a colon. The marker is used to denote proper names (symbol `:r`) and erroneous words (symbol `:q`). The definition of 'erroneous' is strictly limited to typos and errors caused by text conversion – substandard words, dialectical words and frowned-upon expressions are not tagged as erroneous (unless they contain a typo). The symbols can be combined for an erroneous proper name (`:rq`). In the following example, the surname *Kirscher*/`SUfs7:rq` should have been *Kirschner*/`SUfs7:r`.

(1) *duetu* *s* *Janou* *Kirscher*
`SSis2` `Eu7` `SSfs7:r` `SUfs7:rq`
'a duet with Jana Kirsch[n]er'

## 5 MAIN GRAMMATICAL CATEGORIES

### 5.1 *Paradigm*

Slovak language exhibits certain discrepancy between morphological and syntactic behaviour of words (a behaviour shared with other inflected languages). This is reflected in various ways in traditional grammar descriptions, usually by classifying the word to be of a part of speech category corresponding to its morphological class, but acknowledging that the word "behaves as if it were of a different category". Such an ambiguity in description does not have a place in de-

[3] In the following, we count and describe the positions starting with 1, as is customary in many human languages.

signing a morphological tagset, unless we introduce a special category for such ambiguous words, which is something that we were trying to avoid. We introduced the 'paradigm' category, which describes the morphological (inflectional) behaviour of the word. It is in fact a conflation of two different ideas – the inflectional pattern of a different part of speech present in another part of speech, but also describes several other, non-mainstream inflectional patterns.

The paradigm category is specified for nouns, adjectives, pronouns and numerals. The symbol is equal to that of the part of speech category the paradigm follows, with some additional types.

We recognise the following paradigms:

Substantive (symbol `S`) – used for nouns, pronouns and numerals.

Adjectival (symbol `A`) – used for nouns, adjectives, pronouns and numerals.

Pronominal (symbol `P`) – used for pronouns.

Numeral (symbol `N`) – used for numerals.

Adverbial (symbol `D`) – used for pronouns and numerals.

Mixed (symbol `F`) – used for nouns, adjectives, pronouns and numerals. This paradigm is used for words that do not clearly follow one inflectional pattern but instead exhibit features from two or more morphological parts of speech.

Incomplete (symbol `U`) – used for nouns, adjectives, pronouns and numerals. This is used in a case where the word does not exploit all the morphological inflections, typically an uninflected noun or an adjective, 3rd person possessive pronouns and (some) cardinal numerals.[4]

### 5.2 *Grammatical Number*

There are two grammatical numbers in Slovak, singular and plural. Of the old Slavic dual there are only some traces left, but they are not in contrast with singular and plural (unlike e.g. in Czech, where the dual still manifests itself in several nouns – body parts – in the instrumental). We use `s` as the symbol for the singular and `p` for the plural; there are no provisions for marking pluralia and singularia tantum.

[4] Not to be confused with a 'partial' (or 'incomplete') paradigm where a part of the paradigm is missing, such as pluralia tantum.

### 5.3 *Grammatical Gender*

In Slovak, 3 traditional genders are recognised, but in our analysis we split the masculine animate and masculine inanimate to get 4 different genders: masculine animate – `m`, masculine inanimate – `i`, feminine – `f`, neuter – `n`. There are two more 'genders' marked in the tagset, general – `h` and undefined – `o`. These are used as a conflation of other genders in cases where disambiguating them would be impractical or directly impossible. Personal pronouns use the `h` (general) symbol for everything except the third person ones (*on, ona, ono, oni, ony*). In the 1st or 2nd person, the pronouns could be reasonably assigned a gender only in the presence of an adjective or a verb in conditional or past tense – in a typical sentence with a verb in the indicative form it is impossible. Verbs use the `h` for the L-participle plural in the first and second person (in agreement with corresponding personal pronouns, which is also marked with the 'general' gender) and the `o` (undefined) for the third person if the verb covers several genders at once – e.g. the following example has the verb *kričali* (yelled) tagged with the undefined gender, because there are two subjects in the sentence – *muž* is masculine, but *žena* is feminine.

(2) *muž* *a* *žena* *na* *seba* *kričali*
`SSms1 O SSfs1 Eu4 PPhs4 VLepco+`
'[the] man and woman yelled at each other'

### 5.4 *Case*

Slovak distinguishes 6 cases, the locative case being obligatorily prepositional and the nominative obligatorily non-prepositional. We fully realise there is no separate vocative case described by traditional grammars in the contemporary system of Slovak language morphology. What we called a "vocative" in this article is in fact a syntactical role of a noun when used for addressing someone, a role that is only sometimes realised morphologically and in most of the cases is identical with the form of the nominative case. The exceptions exist in the case of several nouns (fossilised forms of old Slavic vocative) such as *bože, pane, priateľu, človeče* … (God, Sir, friend, man) and (substandard usage of) some proper names and interpersonal relationship terms – *Zuzi, babi, oci, mami, tati, šéfe* … (Susan, grandma, dad, mum,

dad, boss). If this article were about Russian, we would use the term "new vocative" here (see e.g. Comtet, 1997).

The cases were traditionally numbered (starting with elementary and secondary school syllabi) and Slovak linguistic and general audience is familiar with case numbers. The numbering went 1-nominative, 2-genitive, 3-dative, 4-accusative, 6-locative, 7-instrumental. We decided to retain this numbering in our tagset, so the numbers 1 through 7 reflect these cases (with the number 5 for the vocative).

### 5.5 *Degree of Comparison*

Slovak has three degrees of comparison: positive, comparative and superlative. The degree is defined only for adjectives, participles and adverbs, and we assigned to it the symbols `x` for positive, `y` for comparative and `z` for the superlative, for all these three parts of speech.

## 6 PART OF SPEECH CATEGORIES

### 6.1 *Noun*

The noun tag is of a fixed length of 5 positions:

| Position | Possible values | Description |
|---|---|---|
| 1 | S | part of speech tag |
| 2 | SAFU | paradigm |
| 3 | mifn | gender |
| 4 | sp | number |
| 5 | 1234567 | case |

The `S` paradigm stands for 'normal' nouns with a full, substantive-like morphology. The `A` (adjectival) paradigm stands for substantivised adjectives or participles. These are often distinguished by proper adjectives only by their semantic role and there often exists an identical adjective or a participle as well. Examples include *obžalovaný* (accused, a passive participle of *obžalovať*), *cestujúci* (traveller, an active participle of *cestovať*), *zelený* (a member of the Green movement; adjective when it is a colour term). The `U` paradigm is used for uninflected nouns – the same form in all the cases and numbers, either completely domesticated loanwords like *kupé*/`SUns1`, *finále*/`SUns1`, or loanwords like *whisky*/`SUfs1`, *miss*/`SUfs1`, or several native substantivised short phrases

or words like *skaderuka*/sUms1-*skadenoha*/sUms1. It is also used for letters (of the alphabet) when used as nouns (e.g. as in the sentence *Od*/Eu2 *A*/sUns2 *po*/Eu4 *Z*/sUns4 *./z*). The F paradigm is used for nouns which combine different inflectional paradigms – e.g. *princezná*/sFfs1 inflects like an adjective, with the exception of genitive (*princezien*/sFfp2), locative (*princeznách*/sFfp6) and instrumental (*princeznami*/sFfp7) plural.

Many animal names in Slovak are masculine animate in the singular, but depending on the familiarity and the degree of anthropomorphisation, the plural can be either animate or inanimate. The rule of thumb is: the higher the organism, the more animateness it shows. People are always animate; *pes* (dog) is animate in the singular and can be both in the plural; *jeleň* (deer) is mostly inanimate in the plural (but can be sometimes animate); *bacil* (germ) is inanimate in the singular and plural, but there are cases of animate singular appearing[5], and *stroj* (machine) is inanimate without exception. The animateness is sometimes used as a semantic disambiguator – *android* is mostly animate when it is a humanoid robot, but mostly inanimate when it is an operating system.

This is reflected in the morphological database – there are lexemes that are masculine animate in the singular and masculine inanimate in the plural, or the plural entries have two variants (inanimate and animate). There are even some singular cases, e.g. *knieža* (duke) is neuter, with the exception of nominative singular, which can be also masculine (the form is the same, but the gender governs adjective and verb agreement), so the tags for *knieža* could be both SSms1 and SSns1.

### 6.2 *Adjective*

The adjective tag is of a fixed length of 6 positions:

| Position | Possible values | Description |
|---|---|---|
| 1 | A | part of speech tag |
| 2 | AFU | paradigm |
| 3 | mifn | gender congruence |
| 4 | sp | number congruence |
| 5 | 1234567 | case congruence |
| 6 | xyz | degree of comparison |

[5] Not all of the examples cited are ‘correct’ by official language rules and dictionaries, they however have non-negligible corpus evidence.

The `U` paradigm stays for indeclinable adjectives. These include a rather exceptional case: three fossilised short forms *hoden*/`AUms1x` (worth), *vinen*/`AUms1x` (guilty) and *dlžen*/`AUms1x` (indebted). These forms also have a different syntactical usage from their regular counterparts. occur only in nominative singular and that is the only entry in the database – the regular long forms *dlžný*/`AAms1x`, *vinný*/`AAms1x`, *hodný*/`AAms1x` are separate lexemes (and the short forms have already shifted semantically). Other indeclinable adjectives are e.g. *nanič*, *akurát*[6], *rád*, special form *naj* (superlative prefix, when used standalone in an adjectival function), and many loanwords at various level of domestication – *super, fajn, hurá, bianko, nealko* …

The `F` paradigm marks possessives (which are considered to be adjectives in traditional Slovak grammars) – e.g. *tetin*/`AFis1` *hlas*/`SSis1`, *Sapkowského*/`AFfp1x` *knihy*/`SSfp1`. The gender in the tag agrees with the gender of the possessed; the gender of the possessor is not marked. It is also used for the special adjective *nesvoj*/`AFms1x`, which is morphologically identical with possessives (derived from a possessive pronoun *svoj*/`PFms1`).

In ambiguous cases (where even traditional grammar descriptions admit that a decision cannot be taken unambiguously) (e.g. *nepočujúci*), we sorted the words according to their attested usage in the corpus – the word was classified as an adjective only if there was a significant percentage of its occurences in adjectival positions (i. e. modifying a noun), disregarding intentionally defective or metalanguage usage. Such decisions have been consulted with the Short Dictionary of the Slovak Language (Považaj, 2003), but preferring the actual corpus evidence.

### 6.3 *Pronoun*

The structure of the pronoun tags depends on the pronoun inflectional paradigm (roughly, the tag structure follows that of the corresponding part of speech of the paradigm type).

[6] Note that *nanič* and *akurát* can be adverbs as well, in adverbial constructions.

| Position | Possible values | | | Description |
|---|---|---|---|---|
| 1 | P | P | P | part of speech tag |
| 2 | SAFU | P | F | paradigm |
| 3 | mifn | h | min | gender |
| 4 | sp | sp | s | number |
| 5 | 1234567 | 1234567 | 24 | case |
| 6 | | | g | agglutinated |

The pronouns are split into three subclasses, according to the paradigm position. The table above captures the three possible combinations of values in the 'Possible values' columns and can be viewed as a concise combination of three different tables, one for the `S, A, F, U` paradigms, the second one for the `P` paradigm and the third one for the `F` paradigm (which is longer by one position, the 'agglutinated' value.

The paradigm `A` is used for adjective-like pronouns: *aký, ktorá, inakšie, samý*.

`F` is used for pronouns that do not have clearly separated morphosyntactical paradigm, typically possessives, e.g. *môj, tvoj, svoj, tento, táto, toto*, and basic personal pronouns *ja, ty, my, vy, seba*.

`U` is used for pronouns that do not decline. These are 3rd person possessives *jeho* (his, its), *jej* (hers) and *ich* (theirs), and *koľko, toľko, bárskoľko, hockoľko* ….

The symbol `g` marks agglutinating of preposition and pronoun – in majority of pronoun tags it does not occur and the tag is then 5 characters long. It appears in pronouns like *preňho*/`PFms4g`, *doň*/`PFms2g` (which are fusions of *pre*/`Eu4` *neho*/`PFms4`, *do*/`Eu2` *neho*/`PFns2`. These pronouns are lemmatised as *pre_on*, *do_ono* (i.e. the combination of a preposition and a pronoun, joined by an underscore). The only existing tags with the agglutinating symbols are `PFms2g`, `PFis2g`. `PFns2g`, `PFms4g`, `PFis4g`. `PFns4g` (i.e. only genitive or accusative singular, non-feminine gender). These agglutinations are traditionally described as pronouns, which was the main reason for including them in this category.[7]

The uninflected adverbial pronouns are tagged with the tag `PD`: *ako, tak, prečo, načo*.

[7] The alternative would be to tokenise them as two different tokens; this would however complicate the tokenisation phase.

6.4 *Numeral*

| Position | Possible values | Description |
|---|---|---|
| 1 | N | part of speech tag |
| 2 | SANFU | paradigm |
| 3 | mifn | gender |
| 4 | sp | number |
| 5 | 1234567 | case |

The paradigm follows the morphology of the numeral, but since the morphology reflects the numeral type, it is also useful for determination of the type.

The `N` paradigm describes small cardinal numerals (2, 3, 4). These are inflected and always in the plural. For the numeral 2 all the genders have separate inflections, and for the cardinality of 3 and 4 the masculine animate gender is in contrast to other genders in the nominative and accusative.

The tag `S` is used for other numerals that inflect like nouns – fractions like *tretina*, *štvrtina*, huge cardinals like *milión*, *septilión* and the word *raz* (once).

The `F` paradigm is used for the cardinal number 1. The numeral is inflected for gender, case and number (the plural is used for group numerals).

The `U` paradigm is used for other cardinal numbers (5, 6, 7, …).

The `A` paradigm describes numerals with adjective-like inflection – primarily ordinal numerals, but also several indefinite ones like *mnohý*`/NAms1`.

The tag `ND` (not inflected, without any other grammar categories) is used for adverbial numerals, e.g. *neraz*, *prvýkrát*.

Not only are the inflectional patterns of several classes of these numerals (noun-like, adjective-like) identical to the corresponding parts of speech, but their syntactic behaviour is also equivalent. In this regard, the usage of the `N` tag differs from other parts of speech, because it encodes also their semantic role. This behaviour was retained from the traditional grammars and the description present in the Short Dictionary of Slovak Language[8].

[8] Apart from the word *polovica* ([one] half), which is considered to be described erroneously as a noun in the dictionary

## 6.5 *Verb*

The tag for verbs is probably the most complicated out of the whole tagset. It does not have a fixed length; the length is determined by the second position, which in this case does not mark the paradigm, but the form of the verb.

We do not adhere strictly to established grammar categories, but follow the verbal form instead. This is the reason we do not mark the tense as such.

Each tag is however at least 4 positions long, and these four positions have a fixed meaning. The third position marks aspect – there are three possible values: `d` for the perfective aspect, `j` for the imperfective one and `e` for the ambivalent verbs. The ambivalent aspect actually means the perfective and imperfective verb forms are identical (but e.g. they form the future tense differently, according to their aspect). Since they are identical in their morphology and we follow strictly formal morphological criteria, we do not try to disambiguate them.

The last position marks positiveness/negativeness – we use the plus sign + for positive verbs and the minus sign - for negated ones. The negation of Slovak verbs is formed with the *ne-* prefix (e.g. *kompiluj* will be negated as *nekompiluj*) invariably in all the conjugated forms. There are some verbs that lack the negated form (e.g. *nenávidieť*/`VIe+`, although some corpus evidence exists, it points out to metalanguage usage or puns) and for these the negated tag does not appear. The only exception is the indicative of the verb *byť* (to be), which is negated by a separate particle *nie*, written separately (this is just an orthography quirk). These cases are tokenised as two separate tokens, with the first one tagged as a particle and the second one as a (positive) verb:

(3) *jazyk* *nie* *je* *usporiadaný*
`SSis1` `T` `VKesc+` `Gtis1x`
'language is not ordered'

This does not explicitly contain information about the 'negativeness', but marking it in any other way would introduce other inconsistencies (e.g. marking the negativeness of a morphologically positive verb or marking the particle as a verb).

We describe the tags sorted by their length, going from the longest (the most complicated) to the shortest one.

| Position | Possible values | Description | Position | Possible values | Description |
|---|---|---|---|---|---|
| 1 | V | part of speech tag | 1 | V | part of speech tag |
| 2 | L | form (L-participle) | 2 | KMB | form |
| 3 | dej | aspect | 3 | dej | aspect |
| 4 | sp | number | 4 | sp | number |
| 5 | abc | person | 5 | abc | person |
| 6 | mifnho | gender | 6 | + − | negation |
| 7 | + − | negation | | | |

The L-participle is used to form past tense(s) and conditionals. Sharing some features with participles, it distinguishes number and gender, and these appear in the tag, making it 7 positions long. Two extra genders – `h` and `o` – were described in Section 5.3.

The indicative (tag `K`) is used to form a present tense for imperfective verbs, and a future tense for perfective ones. For ambivalent-aspect verbs, the indicative form can mean either a present or a future tense, depending on the meaning of the verb. The indicative of the verb *byť* is also used to form the past tense (together with the L-participle). We do not distinguish this auxiliary usage of *byť* from the copula.

The imperative (tag `M`) is also marked for number and person. In the singular, only the second person is possible; in the plural, imperatives can have both the second and the first (inclusive) person.

The future (`B`) is mostly used for the future form of the auxiliary verb *byť*, which is used to form the future tense of imperfective verbs (together with their infinitive), and for the simple future of the copula *byť*. It is also used for a small class of verbs of movement, which form the future tense with the prefix *po-*.

| Position | Possible values | Description |
|---|---|---|
| 1 | V | part of speech tag |
| 2 | IH | form |
| 3 | dej | aspect |
| 4 | + − | negation |

The transgressive (symbol H) in Slovak is morphologically derived from the 3rd person plural indicative, usually by adding the suffix *-c*. It has only one form and in contrast to Czech, it is not distinguished either for number or gender, and there is only a present transgressive. The transgressive is marked just with the aspect and negation – *čítajúc*/VHj+, *neuznajúc*/VHd-.

The infinitive (symbol I) have just one form and is also marked only with the aspect and negation.

### 6.6 *Participle*

There are two different classes of participles in Slovak – active and passive (the L-participle has been discussed in the section on Verbs). The participles exhibit a strong adjective-like morphological behaviour, up to being inflected for a degree of comparison. Their classification as verbs or adjectives is a perennial problem in many languages, and either way leads to some unsatisfactory behaviour. In the Russian Multext East tagset (Sharoff *et al.*, 2008), the participles belong to the Verb category and as such have the case attribute – this has been facilitated by the Multext East formal appearance – the 'case' position is always present, it is just left undefined for the verbs, and it is very easy to reuse it for participles. It is not clear if this coincidence was decisive in categorising the participles or not.

On the other hand, in the Czech Multext East tagset (Dimitrova *et al.*, 1998) the participles are not distinguished in any way from adjectives – they have the 'qualificative adjective' attribute.

We consider the participles to be a separate part of speech class, not a declined form of verbs – while definitely possible, this would lead up to some singular categorisation, e.g. verbs with case. The participles are functionally very similar to adjectives, and indeed many an adjective has originated as a participle of which the source verb is no longer in the language. Sometimes the boundary between participles and adjectives is rather unclear – in ambiguous cases, we conformed to the Short Dictionary of the Slovak Language, which is an arbitrary solution, but probably the best one, given the status of the dictionary. For cases not clearly stated in the dictionary we leaned towards the participles if there existed (at least formally) the source verb.

The passive participle is not distinguished for tense, but formally the active participles are separated into present active and past active ones. The present active participle is commonly found in standard Slovak, but the past active participle is dead for all practical reasons in both literary and standard Slovak – there are only 7 occurrences of the form in the manually annotated corpus, 6 of them from the same document (a treatise on liturgic history).

For this reason, we decided not to introduce any special category separating past and present active participles and use the same tags for both of them. However, we differentiate between passive and active participles.

| Position | Possible values | Description |
|---|---|---|
| 1 | G | part of speech tag |
| 2 | kt | type |
| 3 | mifn | gender congruence |
| 4 | sp | number congruence |
| 5 | 1234567 | case congruence |
| 6 | xyz | degree of comparison |

The type of the participle is marked by the second symbol in the tag – `k` for active, `t` for passive ones. The rest of the symbols follows the symbols for the adjective. Participles can also have a degree of comparison, even if the comparatives and superlatives occur rather rarely.

## 6.7 *Adverb*

The tag for an adverb is invariably two letters long:

| Position | Possible values | Description |
|---|---|---|
| 1 | D | part of speech tag |
| 2 | xyz | degree of comparison |

The degree of comparison is always specified, even if neither comparative nor superlative exists for the adverb (e.g. *neveľmi*/Dx). While we could consider marking irrelevant degrees of comparison (as opposed to positive ones), for the sake of consistency we decided to unify these two cases – this also saves us from having to invent excuses for claiming that a given word has irrelevant degree (according to traditional grammars), even if corpus evidence suggests otherwise.

6.8 *Preposition*

| Position | Possible values | Description |
|---|---|---|
| 1 | E | part of speech tag |
| 2 | uv | vocalisation |
| 3 | 23467 | case (valency) |

Some of the prepositions ending with a consonant exhibit vocalisation – a vowel is appended after the preposition ending with a consonant in certain cases, mostly in non-syllabic preposition, if the next word begins with a consonant of the similar class as the last consonant of the preposition, e.g. if both consonants are sibilants, or both consonants are velar stops, or both consonants are alveolar stops – *k/*$_{Eu3}$ *domu/*$_{SSis3}$ (to [the] house); *ku/*$_{Ev3}$ *korpusu/*$_{SSis3}$ (to [the] corpus), or in some other fixed epressions – *bez/*$_{Eu2}$ *strachu/*$_{SSis2}$ (fearless); *bezo/*$_{Ev2}$ *mňa/*$_{PPhs2}$ (without me).

We mark the vocalised prepositions with the symbol v at the second position, the non-vocalised ones are marked with the symbol u. The lemma of the vocalised prepositions is the non-vocalised form. The third position of the tag encodes the case the preposition binds with (nominative and vocative are not present, according to existing grammar theories).

Compound prepositions are analysed as a sequence of constituents, if possible – e.g. *s/*$_{Eu7}$ *ohľadom/*$_{SSis7}$ *na/*$_{Eu4}$ is tagged as a preposition, a noun and a preposition. There is a sizeable amount of fossilised noun or verb forms that have become prepositions, and these are marked as prepositions (*postupom*, *doprostriedku*, *končiac*). There is often a homonymy with adverbised fossilised forms as well. This makes the class of prepositions less closed and unambiguous than we would like.

In Slovak, no preposition that binds the nominative exists – the reason is that nominative is obligatory non-prepositional and the reason for the nominative to be obligatory non-prepositional is that no prepositions binding with the nominative exist. This circular reasoning is generally accepted in traditional linguistic circles, and the loans *à*, *à la* (often domesticated as *á*, *á la* or *a la*) are not considered to be prepositions, but particles instead.

In our tagset, we did not dare to break this tradition, and *à la* will be tagged as two tokens, a residual *à/*$_{Q}$ and the particle *la/*$_{T}$ (see below for the description of the tags).

(4) *šat* *à* *la* *Zajac*
SSis1 Q T SSms1:r
'clothing à la Zajac'

## 6.9 *Other Categories*

**Conditional morpheme** *by* has a special tag Y. This standalone morpheme is used to form conditionals (with the L-participle), but it can also form multiword prepositions or conjunctions (*nie že by*). However, such multiword prepositions are followed by the L-participle and semantically introduce conditional clauses, therefore it is easier to consider the *by* to be part of the conditional in these circumstances as well. We decided to tag the *by* in all these cases with the same tag. The homonymous poetic conjunction *by* (an abbreviation of *aby*) is also tagged by the Y, which is an inconsistency with other conjunctions, but it is justified by the highly poetic (and therefore rather infrequent) nature of the word and the need to keep the ambiguity low.

Since the morpheme fused with some other functional words, the symbol Y is also used as a second symbol in several other part-of-speech tags, to denote the fusion.

**Punctuation** characters have their own one-letter long tag Z. Lemmas of the punctuation characters are 'normalised' – various types of quotes are lemmatised as straight quotes U+0022 QUOTATION MARK, hyphens and dashes as U+002D HYPHEN-MINUS.

**Conjunctions** can be either simple, having the one-letter tag O (*a, aj, alebo, než...*), or they can contain a fused conditional morpheme *by*, with a two-letter tag OY (*aby, keby, akoby, niežeby, žeby, sťaby, čoby, nietoby, nietožeby*).

**Particles** are tagged with the tag *T*. Similar to conjunctions, some of the particles contain fused conditional morpheme *by* (*čoby, kiežby, žeby*) and are tagged with TY.

**Abbreviations** are tagged with W. We do not distinguish between abbreviations and acronyms, and we do not assign any other grammar categories to it (even if the abbreviated words have them), e.g. in

(5) *odborní pracovníci SNK podali obraz*
AAmp1x SSmp1 W VLdpcm+ SSis4
'SNK professionals gave impression'

the *SNK* is an abbreviation, even if it can be thought of as a noun in genitive singular. As an artifact of our tokenisation, if there is a trailing dot, it is not a part of the token, but a separate token with the punctuation tag Z.

The lack of other categories is indeed debatable – e.g. for the noun-like abbreviations it is reasonable to expect them to have cases, numbers and genders. Our decision was based on the resulting simplicity and the evasion of the need to disambiguate uninflected abbreviations for the values.

**Reflexive morphemes** *sa* and *si* are treated in a special way. They can be a part of a verb as a reflexive morpheme; however they are detached from the verb itself and even their position in the sentence can somewhat vary. The situation is complicated by the fact that *sa* and *si* are also (reflexive) pronouns – an abbreviated form of *seba* and *sebe*, and the distinction of a verbal morpheme and a pronoun is very subtle. If there are more of the pronouns/morphemes in the clause, they usually fuse into one, e.g in the sentence *bojím* **si** *priznať pravdu* there are two verbs *bojím* **sa** and *priznať* **si**.

We solved the problem by assigning a special tag R to both *sa* and *si*, regardless of their function. Unrelated uses of *si* as a 2nd person singular of the verb *byť* and the use of *sa* as a (poetic) particle in (fixed) expressions *sem sa*, *hor sa* are marked as a verb and a particle, respectively.

**Interjections** are tagged with J. Accordingly with the traditional grammars, we also mark greetings as interjections (*ahoj/*J, *ahojte/*J, *čau/*J, *čaute/*J …), where the lemma is always identical with the word form.

**Numbers** written as digits are marked with the tag 0 (the digit zero). Both Arabic and Roman numerals are recognised, the lemma is identical to the word form (except for misspellings, where the lemma is normalised to the 'correct' form, e.g. *l984/*0 with a leading letter instead of digit will be lemmatised as *1984*).

**Undefined part of speech (residual)** is a token that cannot have its part of speech determined – the reason is usually that it is a part of a multiword expression that has been tokenised as several separate tokens. It is tagged with the symbol `Q`. Examples include hyphenated compounds (*sociálno/*`Q` *-/*`Z` *ekonomický/*`AAis1`), components of foreign proper names (*New/*`Q` *York/*`SSis4:r`), but also tokens that are not considered standalone words by traditional grammar theories – expressions like *po/*`Q` *anglicky/*`Dx`, *na/*`Q` *modro/*`Dx`, where the whole expression will be considered an adverb.

**Foreign language citation** is reserved for citation elements, i.e. foreign language words that appear to be foreign elements in the text (neither loanwords, nor commonly used proper names). Typically, these are short citations, names of books, movies etc. The symbol for this tag is `%` `U+0025 PERCENT SIGN`.

**Non-word element** is anything that is neither a word nor punctuation. Typically, these are remnants of incorrect conversion (which would not be there in an ideal world), (pseudo)graphical elements, fancy paragraph separators. A simple test deployed in the tagging process is to consider non-word elements tokens that do not belong to a fixed set of common punctuation characters and that do not consist of alphanumeric characters. The symbol for a non-word element is `#` `U+0023 NUMBER SIGN`. A typical example is the copyright sign ©*/*`#`.

## 7 CONCLUSION

We have described the tagset designed and used in the Slovak National Corpus. The tagset is used in a morphological database of Slovak words and in the manually annotated corpus of Slovak language, *r-mak*. The database and the manually annotated corpus are then used to train an automatic morphological tagger *morče* (Votrubec, 2006) developed at the Faculty of Mathematics and Physics, Charles University and used originally to tag the Czech language. Morče is used for automatic lemmatisation and tagging of the whole Slovak National Corpus and other Slovak language corpora and subcorpora. The tagset has become de facto the standard tagset used in automatic morphosyntactic analysis and tagging of Slovak language texts. The complete tagset tables with examples can be found online at `http://korpus.sk/morpho.html`.

## REFERENCES

Vladimír BENKO, Jana HAŠANOVÁ, and Eduard KOSTOLANSKÝ (1998), *Model morfologickej databázy slovenčiny. Počítačové spracovanie jazyka*, Pedagogická fakulta Univerzity Komenského, Bratislava, Slovakia.

Roger COMTET (1997), *Grammaire du russe contemporain*, Presses Universitaires du Mirail.

Łukasz DĘBOWSKI (2001), Tagowanie i dezambiguacja, in *Prace IPI PAN 934*, Instytut Podstaw Informatyki PAN, Warsaw, Poland.

Ludmila DIMITROVA, Tomaž ERJAVEC, Nancy IDE, Heiki Jaan KAALEP, Vladimir PETKEVIČ, and Dan TUFIŞ (1998), Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages, in *Proceedings of the COLING-ACL'98*, pp. 315–319, Montréal, Québec, Canada.

Ladislav DVONČ, Gejza HORÁK, František MIKO, Jozef MISTRÍK, Ján ORAVEC, Jozef RUŽIČKA, and Milan URBANČOK (1966), *Morfológia slovenského jazyka*, Vydavateľstvo Slovenskej akadémie vied, Bratislava, Slovakia.

Sašo DŽEROSKI, Tomaž ERJAVEC, and Jakub ZAVREL (2000), Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagset, in *Proceedings of the Second International Conference on Language Resources and Evaluation*, pp. 1099–1044, ELRA, Paris, France.

Radovan GARABÍK (2006), Slovak morphology analyzer based on Levenshtein edit operations, in *Proceedings of the WIKT'06 conference*, pp. 2–5, Institute of Informatics SAS, Bratislava, Slovakia.

Radovan GARABÍK (2011), Slovak MULTEXT-East Morphology tagset, *Jazykovedný časopis*, (1):19–39.

Radovan GARABÍK, Lucia GIANITSOVÁ, Alexander HORÁK, and Mária ŠIMKOVÁ (2004), Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu, URL `http://korpus.sk/attachments/publications/2004-garabik-gianitsova-horak-simkova-tokenizacia.pdf`, Internal documentation.

Radovan GARABÍK, Daniela MAJCHRÁKOVÁ, and Ludmila DIMITROVA (2009), Comparing Bulgarian and Slovak Multext-East morphology tagset, in *Organization and Development of Digital Lexical Resources*, pp. 38–46, Dovira Publishing House, Kyiv, Ukraine.

Jan HAJIČ (2004), *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, Karolinum, Charles Univeristy Press, Prague, Czech Republic.

Jan HAJIČ (2000), Morphological Tagging: Data vs. Dictionaries, in *Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference*, pp. 94–101.

Jan HAJIČ and Barbora VIDOVÁ-HLADKÁ (1997), Morfologické značkování korpusu českých textů stochastickou metodou, 4(58):288–304.

Matej POVAŽAJ, editor (2003), *Krátky slovník slovenského jazyka. 4., doplnené a upravené vydanie*, Veda, Bratislava, Slovakia.

Emil PÁLEŠ (1994), *SAPFO. Parafrázovač slovenčiny. Počítačový nástroj na modelovanie v jazykovede*, Veda, Bratislava, Slovakia.

Radek SEDLÁČEK (2001), A new Czech morphological analyser ajka, in *Proceedings of the TSD, Czech Republic*, pp. 100–107, Springer Verlag.

Serge SHAROFF, Mikhail KOPOTEV, Tomaz ERJAVEC, Anna FELDMAN, and Dagmar DIVJAK (2008), Designing and Evaluating a Russian Tagset, in Nicoletta CALZOLARI, Khalid CHOUKRI, Bente MAEGAARD, Joseph MARIANI, Jan ODJIK, Stelios PIPERIDIS, and Daniel TAPIAS, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco, URL `http://www.lrec-conf.org/proceedings/lrec2008/`.

Pavel ŠMERK (2010), A New Data Format for Czech Morphological Analysis, in *Proccedings of Recent Advances in Slavonic Natural Language Processing*, pp. 3–8, Tribun EU, Karlova Studánka, Czech Republic, URL `http://www.fi.muni.cz/sojka/download/raslan2010/raslan10.pdf`.

Jan VOTRUBEC (2006), Morphological Tagging Based on Averaged Perceptron, in *WDS'06 Proceedings of Contributed Papers*, pp. 191–195, Matfyzpress, Charles University, Praha, Czech Republic.